View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilPapers

Hypotheses that Attribute False Beliefs − a Two-Part Epistemology (Darwin+Akaike)

William Roche* and Elliott Sober†

* Department of Philosophy, Texas Christian University, Fort Worth, TX, USA, e-mail: w.roche@tcu.edu

† Department of Philosophy, University of Wisconsin, Madison, Madison, WI, USA, e-mail: ersober@wisc.edu

Abstract: Is there some general reason to expect organisms that have beliefs to have false beliefs? And after you observe that an organism occasionally occupies a given neural state that you think encodes a perceptual belief, how do you evaluate hypotheses about the semantic content that that state has, where some of those hypotheses attribute beliefs that are sometimes false while others attribute beliefs that are always true? To address the first of these questions, we discuss evolution by natural selection and show how organisms that are risk-prone in the beliefs they form can be fitter than organisms that are risk-free. To address the second question, we discuss a problem that is widely recognized in statistics – the problem of over-fitting – and one influential device for addressing that problem, the Akaike Information Criterion (AIC). We then use AIC to solve epistemological versions of the disjunction and distality problems, which are two key problems concerning what it is for a belief state to have one semantic content rather than another.

Keywords: Akaike Information Criterion, belief, disjunction problem, distality problem, error, expected utility, evolution by natural selection, misrepresentation, predictive accuracy.

1 Introduction

A central problem facing naturalistic accounts of belief content[1] is to show how beliefs can be false. The possibility of error is widely supposed to distinguish belief from non-mentalistic items that have "natural meaning." Grice (1957) coined this phrase, and described how it applies to

[1] By a naturalistic account of belief content we mean a theory that characterizes necessary and sufficient conditions for a belief's having a given semantic content, where the conditions are given exclusively in terms of nonintentional concepts deployed in natural sciences. Such theories are reductive; see McLaughlin (1987) and Shapiro (1997) for discussion. They may use concepts from information theory and statistics, causal concepts, and functional concepts like adaptation from biology.

token events. He gives the example of "those spots mean measles" and says "*x meant that p* and *x means that p* entail *p*" (Grice 1957, p. 377, italics original). The spots don't mean measles if the person in question has the spots but lacks the measles. Natural meaning, in short, is factive. Beliefs are not like this.

The problem of explaining how misrepresentation is possible can be brought into focus by considering the much-discussed example of a frog that catches flies with its tongue.[2] Take some frog F, and suppose that a fly flies by, this causes F to go into neural state n, and the latter, in turn, causes F to shoot out its tongue and catch the fly. Plausibly, n's content is "that's food" or "that's a fly" as opposed to "that's a fly or a beebee". If so, then when a beebee flies by and the frog shoots out its tongue and catches the beebee, which it then swallows, n's content is false, and this is a case of misrepresentation. Any adequate naturalistic account of belief content needs to allow for cases like this. Fodor's (1984) infamous "disjunction problem" can be understood as the problem of meeting this adequacy condition. He argues that various "Wisconsin-style" naturalistic theories fail to do so, and thus are inadequate.[3]

In what follows, we begin with the disjunction problem in spite of the fact that the "disjunction" in the disjunction problem is something of a distraction if the goal is to understand how misrepresentation can arise. The above story about the frog could be told where the puzzle is to say why "that's a fly" is the meaning of the belief state rather than "that's a dark object." The point is that the first belief content is more liable to be wrong than the second (if all flies are dark objects, but not conversely). The disjunctive formulation will be distracting if you think that having the belief "that's a fly or a beebee" requires one to have the concept *fly* and the concept *beebee*. It therefore seems weird to attribute this disjunctive belief to the frog if the frog can't distinguish flies from beebees. So the disjunctive hypothesis sounds absurd from the start, thus reducing to absurdity any theory that entails it. Even so, the broader problem of explaining the possibility of false beliefs remains. That said, we'll stick with the disjunctive formulation of the problem (at least initially) because it is vivid and has been so widely discussed.[4]

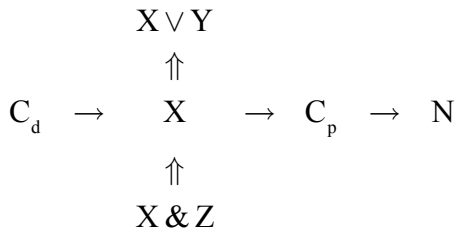
---

[2] The science behind this work was pioneered by Jerome Lettvin, et al. (1959).

[3] Fodor's (1984) targets in particular are the causal/teleological and informational theories of meaning developed by Stampe (1977) and Dretske (1981). We think that Stampe and Dretske anticipated the disjunction problem, and that their theories are not subject to Fodor's objection (Gibson 1996).

[4] It doesn't matter for our purposes whether it's plausible that the frog believes "that's a fly" or "that's food" (for example) when it shoots out its tongue and catches a fly. It isn't our goal to settle exactly what the frog believes. See Neander (2017, p. 262, endnote 2 in Ch. 5) for references on what philosophers have said about this.

Figure 1



The disjunction problem has a twin, as illustrated in Figure 1, which represents two chains. Let "n" denote a neural state, and "N" denote a proposition to the effect that a given organism is in state n. The horizontal chain is causal; the distal cause C_d causes X, X causes the proximal cause C_p, and C_p causes N.[5] If cause must precede effect, this chain has earlier events to the left and later events to the right. The vertical chain is logical; X&Z logically entails X but not conversely (suppose), and X logically entails X∨Y but not conversely (suppose). Take some case where it seems plausible to say that n's content is X. The disjunction problem concerns the vertical chain. An adequate naturalistic theory of belief content needs to allow for and explain cases where n's content is X and not the logically weaker proposition X∨Y or the logically stronger proposition X&Z.[6] The distality problem concerns the horizontal chain. An adequate naturalistic theory of belief content needs to also allow for and explain cases where n's content X and not the more causally distal proposition C_d or the more causally proximal proposition C_p.[7] The distality problem is diachronic; the disjunction problem is synchronic.

It might help to reformulate the two problems as questions relative to a datum. The datum is:

> DATUM: There are cases where organism O is in neural state n and n's content is X as opposed to the logically weaker proposition X∨Y, the logically stronger proposition X&Z, the more causally distal proposition C_d, and the more causally proximal proposition C_p.

The questions are:

[5] We're being a bit loose here in speaking of *propositions* as causes and effects. We really have in mind the *events* described in the various propositions (as causes and effects).

[6] We use "n means P," "n has P as its content," and "n's content is P" interchangeably.

[7] Neander (2017, Ch. 9), referencing Sterelny (1990), calls the first problem "the vertical problem" and calls the second problem "the horizontal problem."

$DISJ_1$: What makes it the case that n's content is X as opposed to the logically weaker proposition X∨Y or the logically stronger proposition X&Z?

$DIST_1$: What makes it the case that n's content is X as opposed to the more causally distal proposition C_d, and the more causally proximal proposition C_p?

Each of these questions is distinct from each of the following:

$DISJ_2$: What counts as evidence for the hypothesis that n's content is X as opposed to the logically weaker proposition X∨Y and the logically stronger proposition X&Z?

$DIST_2$: What counts as evidence for the hypothesis that n's content is X as opposed to the more causally distal proposition C_d and the more causally proximal proposition C_p?

We interpret the "makes it the case" locution in type-1 questions to be metaphysical and constitutive (not causal). Questions of type-2 are epistemological. Our focus in what follows will be on questions similar to (though not identical with) $DISJ_2$ and $DIST_2$. We thus will leave $DISJ_1$ and $DIST_1$ behind.[8]

It's important to distinguish between the question of whether an organism has beliefs and the question of whether an organism that has beliefs has false beliefs. Our target questions in what follows *presuppose* that the organisms under discussion have beliefs.

The remainder of this paper is organized as follows. In Sections 2 and 3, we address our first target question. We do this by considering the fact that organisms evolve by natural selection, and by identifying circumstances in which natural selection leads organisms to form risky beliefs as opposed to beliefs that are risk-free. In Sections 4 and 5, we turn to our second and third target questions. In Section 4, we describe how the Akaike Information Criterion explains when and why simple models that postulate errors often should be expected to be more predictively accurate than complex models that do not. In Section 5, we use that idea to solve epistemological versions of the disjunction and distality problems. In Section 6, we close by providing a brief summary, and relating our findings to Davidson's much-discussed "principle of charity."

2 When selection favors organisms that have lower error probabilities

Suppose that organisms do better with respect to surviving and reproducing if they believe a true proposition rather than disbelieve it. From this it might seem to follow that natural selection will favor belief acquisition mechanisms that are more reliable (a term that requires clarification) over ones that are less. The premise of this argument has exceptions, of course. For example, believing a fairy tale can make you happy, and happiness might make you better at surviving and

[8] See Adams and Aizawa (2017) and Neander (2018) for further discussion and references.

reproducing. And your fitness is often unaffected by whether you believe a given proposition or its negation. But let's set such exceptions aside. Return with us now to those thrilling days of yesteryear when our ancestors had cognitive capacities that were much more limited than our own. And let's suppose that they didn't have distinct beliefs and desires. Beliefs, or something like them, were enough to prod these organisms to action.[9]

Table 1: Pay-offs and probabilities in each of four situations		
	States of the world (and their probabilities)	
	P is true (p)	**P is false (1-p)**
Believe P	$x+b_1$, $1-e_1$	y, e_2
Believe not-P	x, e_1	$y+b_2$, $1-e_2$

The simple argument stated at the start of the previous paragraph can now be given a formal representation. Each of the four cells in Table 1 represents a utility followed by a probability. Consider the upper-left cell. Proposition P is true with probability p. If P is true, the organism has a probability of $1-e_1$ of believing P. If it does so believe, its payoff is $x+b_1$. The other three cells follow suit. The two b's in the table are positive (they are benefits), so this table embodies the assumption that if proposition X is true, an organism does better if it believes X than if it believes not-X.

Notice that e_1 and e_2 are error probabilities. The smaller the error probabilities, the more reliable a belief-formation device is. Field (1990, p. 106) usefully distinguishes head-to-world reliability from world-to-head reliability. The former has to do with low values of Pr(not-P | S believes P) and Pr(P | S believes not-P). The latter concerns low values for Pr(S believes P | not-P) and Pr(S believes not-P | P). In Table 1, low error probabilities represent high degrees of world-to-head reliability. Error probabilities, like the probabilities of states of the world, are to be interpreted objectively; they don't need to represent anyone's credences.

[9] In Gendler's (2008) terminology, these organisms had "aliefs," not beliefs. Allowing for a division of labor between beliefs and desires introduces complications that we won't discuss; see Stephens (2001) and Sterelny (2003).

Let "BFS" describe the *belief-forming strategy* captured in Table 1. The expected utility of BFS is given by:

EU(BFS) =

$p[(x+b_1)(1-e_1) + x(e_1)] + (1-p)[y(e_2) + (y+b_2)(1-e_2)] =$

$p[x + b_1(1-e_1)] + (1-p)[y + b_2(1-e_2)]$

If the values of probability p and utilities x, b_1, y, and b_2 are the same for all the organisms in the population, then the only way one organism can do better than another is by having lower error probabilities e_1 and e_2. The fittest possible organism in this setting is one whose error probabilities are zero.

Why, then, do organisms have false beliefs? One possibility is simply that there are physical limits on how reliable belief formation devices can be. Organisms have false beliefs for the same reason that zebras lack machine guns with which to repel lion attacks (Krebs and Davies 1981). Natural selection is limited in what it can achieve by the range of variation actually present in the population.

Our first target epistemological question concerns whether there could be other reasons why organisms have false beliefs. In particular, consider:

> Q_1: Is there some general reason, apart from the fact that perfect reliability may be impossible, to expect organisms that have beliefs to have false beliefs?

Although the preceding discussion might suggest that the answer is *no*, we argue in the next section that the answer is *yes*.

3 When selection favors organisms that have higher error probabilities

In the argument presented in the previous section, we assumed that the organisms in the population all form beliefs about whether P is true. They all face the same problem. Now let's drop that assumption and suppose instead that different organisms work with different propositions. For example, consider two frogs. One never goes wrong in what it believes, while the other sticks its neck out. Under what circumstances will the risk-taking frog do better than the one that is immune to error? If the two frogs were assessing the same set of propositions and deciding what to believe, the error-free frog would be better off (as just explained). However, the two frogs we're now considering work with different sets. The error-free frog decides whether to believe FLY-or-BB or believe not-FLY&not-BB, where "FLY" denotes the proposition that a fly is present, and "BB" denotes the proposition that a beebee is present. The risk-taking frog decides whether to believe FLY, believe BB, or believe not-FLY&not-BB.

Table 2: The Belief-Forming Strategy SAFE			
	States of the world (and their unconditional probabilities)		
	FLY (s_1)	**BB (s_2)**	**not-FLY&not-BB (s_3)**
Believe FLY-or-BB	x_1+b_1, 1	x_2-c, 1	x_3, 0
Believe not-FLY&not-BB	x_1, 0	x_2, 0	x_3+b_2, 1

Table 2 describes the strategy followed by the risk-free frog. There are three mutually exclusive and collectively exhaustive states of the world; the probabilities for these states are s_1, s_2, and s_3, which sum to one. When a fly or a beebee is present, the frog believes FLY-or-BB. When neither a fly nor a beebee is present, it believes not-FLY&not-BB. This frog never has a false belief, but there is a down-side to its risk-free strategy. When the organism believes FLY-or-BB, it always produces the same behavior. Suppose that behavior is *eating*. Whenever a fly is present, the frog eats the fly and gains a benefit b_1; when a beebee is present, the frog eats the beebee and pays cost c. The expected utility of the risk-free strategy is EU(**SAFE**) = $s_1(x_1+b_1) + s_2(x_2-c) + s_3(x_3+b_2)$.[10]

Table 3: The Belief-Forming Strategy RISKY			
	States of the world (and their unconditional probabilities)		
	FLY (s_1)	**BB (s_2)**	**not-FLY&not-BB (s_3)**
Believe FLY	x_1+b_1, 1-e_1	x_2-c, e_2	x_3, 0
Believe BB	x_1, e_1	x_2, 1-e_2	x_3, 0

[10] Notice that we are here discussing the expected utilities of *strategies* for forming beliefs, which are distinct from the expected utilities of a given *belief*. This corresponds to the familiar distinction in evolutionary game theory between behavioral strategies and behaviors. For example, in the iterated prisoners' dilemma, each act of defection and cooperation has a fitness pay-off, but the fitness of a strategy like tit-for-tat involves a different type of calculation (Axelrod 1980). See McKay and Dennett (2009) for discussion of the fitness consequences of specific beliefs. Notice also that we are not assuming the frogs can literally *choose* which belief-forming strategy to follow.

Believe not-FLY&not-BB	x_4, 0	x_5, 0	x_3+b_2, 1

Now consider the risk-taking frog depicted in Table 3. When that frog faces something that is neither a fly nor a beebee, it unfailingly believes that the object at hand is neither a fly nor a beebee. It makes no mistakes in such cases. However, if a fly is present, the frog has a probability of $1-e_1$ of believing that a fly is present, and a probability of e_1 of thinking that a beebee is present. If a beebee is present, the frog has a probability of $1-e_2$ of believing that a beebee is present, and a probability of e_2 of believing that a fly is present. Here, as before, e_1 and e_2 are error probabilities. If a fly is present and the frog believes FLY, the frog eats the fly and gains benefit b_1; if a beebee is present and the frog believes FLY, it eats the beebee and pays cost c. And so on. The expected utility of this risk-taking strategy is EU(**RISKY**) = $s_1(1-e_1)(x_1+b_1) + s_1(e_1)x_1 + s_2(e_2)(x_2-c) + s_2(1-e_2)(x_2) + s_3(x_3+b_2)$.

The two frogs just described have something in common. In each table, given any state of the world, the fly does better by having a true belief about that state than by having a false one. However, a bit of algebra reveals that the expected utilities of the two strategies can differ:

(*) EU(**RISKY**) > EU(**SAFE**) precisely when $c(s_2)(1-e_2) > b_1(s_1)e_1$.

The following three conditions jointly suffice for the right-hand inequality in (*) to be true:

(i) $c > b_1$: the cost of eating a beebee exceeds the benefit of eating a fly.

(ii) $s_2 > s_1$: Pr(BB) > Pr(FLY).

(iii) $1-e_2 > e_1$: Pr(believe BB | BB) > Pr(believe BB | FLY).

The third of these conditions concerns world-to-head probabilities, and says in effect that the frog is at least minimally world-to-head reliable in distinguishing between beebees and flies. And conditions (ii) and (iii) together entail that the frog is at least minimally head-to-world reliable as well, in that Pr(BB | believe BB) > Pr(FLY | believe BB).[11] But note: if e_1 and e_2 are greater than 0 and conditions (i)-(iii) are satisfied, then **RISKY** is inferior to **SAFE** both in head-to-world reliability and in world-to-head reliability even though **RISKY** has the higher expected utility.[12]

[11] This follows from the odds version of Bayes's theorem: For any E, H_1, and H_2, $Pr(H_1 | E)/Pr(H_2 | E) = Pr(H_1)/Pr(H_2) \times Pr(E | H_1)/Pr(E | H_2)$. This says that the ratio of posterior probabilities equals the ratio of prior probabilities times the ratio of likelihoods.

[12] If a frog follows **RISKY**, then it never suspends judgment on FLY, on BB, or on not-FLY&not-BB. However, this isn't essential. It's straightforward to modify **RISKY** so that suspension of judgment is an option and it's still the case that EU(**RISKY**) can be greater than EU(**SAFE**).

It follows that:

> R_1: There are belief-forming strategies BFS_1 and BFS_2 such that (i) $EU(BFS_1) > EU(BFS_2)$, (ii) if P is true, an organism does better if it believes P than it does if it believes not-P, (iii) BFS_1 is less world-to-head reliable than BFS_2, and (iv) BFS_1 is less head-to-world reliable than BFS_2.

This result is due to the fact that different belief-forming strategies can involve different partitions of propositions, and some partitions lead to more error while at the same time permitting finer discriminations to be drawn that make an adaptive difference in the organism's behavior.

Although conditions (i), (ii), and (iii) are jointly sufficient for the right-hand inequality in (*) to be true, they are not individually necessary. Consider conditiom (iii), for example. Risk-taking frogs can do better than risk-averse frogs even when condition (iii) fails to hold. Risk-free frogs eat flies 100% of the time when flies are present. There is no way for risk-taking frogs to do better than this when the object at hand is a fly. But risk-free frogs also eat beebees 100% of the time when beebees are present. Here risk-taking frogs have an opening. If they eat beebees less than 100% of the time when beebees are present, they are in that respect doing better than their risk-free competitors. They can gain that edge even if condition (iii) doesn't hold because e_1 and e_2 are each greater than or equal to 0.5. If beebees are sufficiently more common than flies and the cost of eating a beebee is sufficiently greater than the benefit of eating a fly, then risk-taking frogs win the competition mainly because they decline to eat beebees, not because they are quick to eat flies. If, for example, $s_2 = 5/8 > 1/8 = s_1$, $c = 80 > 2 = b_1$, and condition (iii) doesn't hold because $0.5 \leq e_1 = e_2 < 200/201$, then $c(s_2)(1\text{-}e_2) > b_1(s_1)e_1$.[13]

This numerical example is interesting in at least two respects. First, the error probabilities e_1 and e_2 can be extremely high. Risk-taking frogs can almost always believe BB when FLY is true, and almost always believe FLY when BB is true. This would be extremely low world-to-head reliability. Second, given the values stipulated for s_1 and s_2, and given that risk-taking frogs never believe not-FLY&not-BB when FLY is true or BB is true, it follows that the values for Pr(FLY | believe FLY) and Pr(BB | believe BB) can be very close to 0. This would be very low head-to-world reliability. Hence EU(**RISKY**) can be greater than EU(**SAFE**) even when risk-taking frogs score poorly—indeed extremely poorly—*both* in terms of world-to-head reliability and in terms of head-to-world reliability.

We assumed above that the risk-free frog always performs the same behavior when a fly or a beebee is present – *eating*. A different criterion would have ensued had we assumed that that single behavior is *not eating*, but the take-home lesson would have been the same: the risk-free

[13] It might be that following **RISKY** requires that both e_1 and e_2 be less than 0.5. We aren't claiming otherwise. Our claim is a conditional. It's worth noting, though, that any case where both e_1 and e_2 are less than 0.5 is a case where $1\text{-}e_2 > 0.5 > e_1$ and thus condition (iii) holds.

strategy has the virtue of never leading to a false belief, but it also has the defect of treating flies and beebees in the same way. Here we see an analogue of James's (1897, p. 27) observation that skepticism has the virtue of avoiding false beliefs but the defect of missing out on true ones.[14] Criterion (*) describes how a belief-forming strategy that yields false beliefs may have a selective advantage over a belief-forming strategy that never does. A strategy that increases the probability of error incurs a cost, but the strategy can provide a compensating benefit in the form of heightened discrimination (with respect to the organism's treatment of flies and beebees).

It's worth noting that R_1 is distinct from each of the following alternative results:

R_2: There are belief-formation strategies BFS_1 and BFS_2 such that (i) $EU(BFS_1) > EU(BFS_2)$, (ii) BFS_1 is less head-to-world reliable than BFS_2, and (iii) BFS_1 is more world-to-head reliable than BFS_2 (see Godfrey-Smith 1991).

R_3: There are belief-formation strategies BFS_1 and BFS_2 such that (i) $EU(BFS_1) > EU(BFS_2)$, (ii) BFS_1 is an "*a priori* prejudice" strategy (decide what to believe without new observational evidence), and (iii) BFS_2 is a "learning" strategy (obtain new observational evidence before deciding what to believe) (see Sober 1994).

R_4: There are belief-formation strategies BFS_1 and BFS_2 such that (i) BFS_1 is superior to BFS_2 in terms of overall information yield, (ii) there is information-theoretic equivocation in the information channel involving BFS_1, and (iii) there is no such equivocation in the information channel involving BFS_2 (see Clark 1993).

A key difference between R_1, on the one hand, and R_2 and R_3, on the other, is that the belief-forming strategies at issue in R_2 and R_3 assume the same partion of propositions. This difference doesn't carry over to R_1 and R_4, but R_1 concerns expected utility whereas R_4 concerns overall information yield.[15]

R_1 should be understood as a *possibility* claim. We haven't argued that any actual organism uses belief-forming strategies of the sort described by R_1. It seems plausible, though, that the sufficient condition (i)-(iii) for EU(**RISKY**) > EU(**SAFE**) is often satisfied in nature, though here we're not thinking about frogs, beebees, and flies, but about a more general pattern. Many organisms are able to distinguish things that are nutritious from things that are not, and

[14] Our assumption that both frogs are such that Pr(believe not-FLY&not-BB | not-FLY&not-BB) = 1 is unrealistic, but it is inessential for our purposes. Our main point here goes well beyond this idealization, in that a criterion much like (*) can be stated where all belief states are capable of being misrepresentations.

[15] Clark (1993, p. 307) notes that it's plausible that greater overall information yield could come with a "selective advantage," but he doesn't develop the point. Our argument for R_1 shows that there's no need to appeal to information theory in order to show the potential value of working with "riskier" partitions of propositions.

live in environments in which things that are bad to eat are more common than things that are good, where the harm that comes from eating something that is bad is, on average, greater than the benefit that comes from eating something that is good.

In addition, there are other sufficient conditions that become visible when the (*) criterion is rewritten in terms of ratios:

(**) EU(**RISKY**) > EU(**SAFE**) if and onl if $c/b > (s_1/s_2)(e_1/(1-e_2))$.

This rewrite describes a relationship between the cost/benefit ratio and ratios of probabilities. It's clear from this reformulation that each of the following conditions also suffices for EU(**RISKY**) > EU(**SAFE**):

- $c >> b_1$, s_2/s_1 is approximately equal to 1, and $e_1/(1-e_2)$ is approximately equal to 1.
- $s_2 >> s_1$, c/b_1 is approximately equal to 1, and $e_1/(1-e_2)$ is approximately equal to 1.
- $1-e_2 >> e_1$, c/b_1 is approximately equal to 1, and s_2/s_1 is approximately equal to 1.

There are lots of ways, then, for the right-hand side of (**) to hold and thus lots of ways for the right-hand side of (*) to hold.

Return now to Q_1 (which we repeat for convenience):

Q_1: Is there some general reason, apart from the fact that perfect reliability may be impossible, to expect organisms that have beliefs to have false beliefs?

The answer, we suggest, is affirmative. Given the evolutionary considerations just described, the hypothesis that present-day organisms with beliefs *sometimes* have false beliefs is more probable than the hypothesis that they *never* have false beliefs (and so the former hypothesis has a probability greater than 0.5).[16]

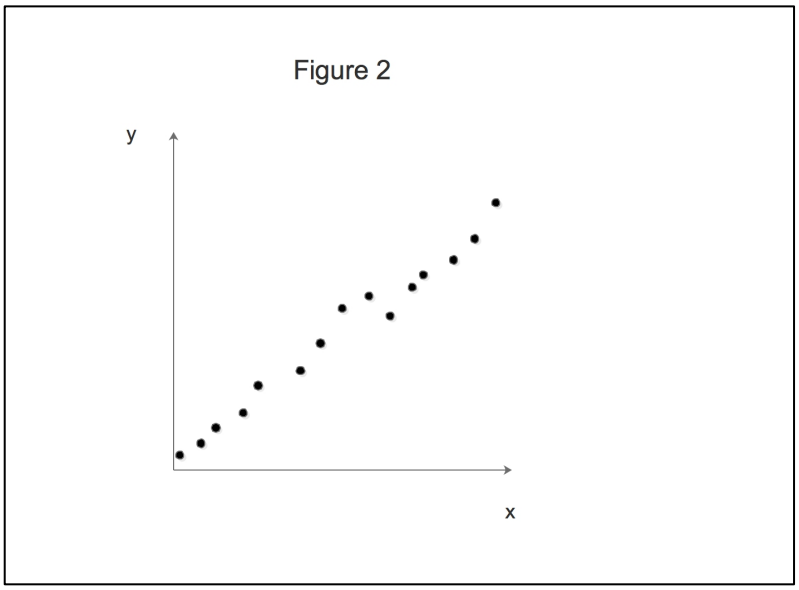
Notice that this broad conclusion leaves it open which minded organisms have which false beliefs. That more specific question requires attention to the specifics of the organism in question, which is the topic to which we now turn, leaving general evolutionary considerations (mostly) behind.

4 AIC, Error, and Predictive Accuracy

We have assumed so far that if proposition P is true, an organism does better by believing P than by believing not-P. That was true in Tables 1, 2, and 3. We set happy-making fairy tales aside, and will continue to do so here. The novelty we want to consider now involves predictive

[16] We are not addressing the question of why natural selection should cause organisms to evolve the capacity to form beliefs (on which see Godfrey-Smith 1996 and Sober 1997). Rather, we are talking about the kind of belief-formation strategies that natural selection will cause organisms to evolve if they have beliefs.

accuracy. It might seem that the road to accurate predictions is to find true beliefs, but this roadmap is mistaken in a kind of situation that matters to science. Scientific models often include idealizations, and idealizations are false. When scientists test idealized models against each other, the question is not which is true or probably true. None of them is. The testing concerns which models can be expected to be more predicively accurate than which others.



Consider the curve-fitting problem depicted in Figure 2. Suppose that the variables x and y are both nonmental; for example, they might be the quantity of July rainfall and the height of August corn plants in a given locale. No straight line fits these data points perfectly, but there is a straight line that fits them pretty well. In contrast, since there are 15 data points, the 14-degree polynomial

(POLY-14) $y = a_0 + a_1x + a_2x^2 + \ldots + a_{14}x^{14}$

fits them perfectly. This polynomial has fifteen adjustable parameters (a_0, a_1, …, a_{14}). A linear model that allows that error is possible has the form

(LIN) $y = b + mx + e$

and thus has three adjustable parameters (b, m, e). The error term e in LIN describes a symmetric bell-shaped distribution for the possible y values that may arise for a given x value.[17] The data allow you to estimate the values of the adjustable parameters.[18]

If you have 15 observations, scientists will usually tell you to shun POLY-14 because it is too complex. The perfect fit this model achieves is not a badge of honor; rather, it is a scarlet letter, marking the ill-gotten booty obtained by egregious over-fitting. Scientists will not be surprised if this model does a poor job of predicting new data when fitted to the data at hand.

The problem of over-fitting is widely recognized in statistics, and there are a variety of strategies that statisticians recommend and that scientists deploy to cope with it. The one we'll discuss here is the Akaike Information Criterion (AIC). The points we will make about attributing errors to organisms do not depend on AIC's being the only or the best tool for the job.

AIC provides a mathematical explanation for why LIN should be expected to be more predictively accurate than POLY-14, given the data in Figure 2; this explanation goes beyond the fact that our past experience reveals that simpler models often have better predictive track records. Akaike (1973) proved that AIC provides an unbiased estimate of how well a model will predict new data when fitted to old (Forster and Sober 1994; Burnham and Anderson 2002; Sober 2015). Model M's AIC score is defined as follows:

$$AIC(M) = \log[Pr(data \mid f(M))] - k.$$[19]

The logarithm here is the natural logarithm. When the adjustable parameters in M are replaced by their maximum likelihood estimates, the result is the fitted model f(M). The first term in the AIC score – $\log[Pr(data \mid f(M))]$ – represents how well f(M) fits the data at hand. The second term – k – represents the model's complexity, as measured by the number of adjustable parameters it contains. The minus sign indicates that AIC penalizes models for their complexity. More complex models generally fit the data better than simpler models do, but AIC penalizes complex models more than it penalizes simpler ones. Given that Pr(data | f(POLY-14)) equals 1, it follows that:

[17] The "error term" in LIN does not mean that nature aims to fall on a straight line but then makes a mistake. That was how the "law of errors" was first understood, but that is now ancient history. For discussion of that history, and references, see Sober (1980).

[18] We are assuming for simplicity that 15 data points is enough for legitimate model selection, but this is dubious. Burnham and Anderson (2002, p. 50), for instance, recommend that you have at least 40 data points for each of the parameters in a model you wish to consider.

[19] Statisticians follow Akaike (1973) and define the AIC score in terms of predictive *in*accuracy, which is -1 times the quantity stated here.

$$\begin{aligned}\text{AIC}(\text{POLY-14}) &= \log[1] - 15 \\ &= 0 - 15 \\ &= -15\end{aligned}$$

If, say, Pr(data | f(LIN)) equals 2/3, then:

$$\begin{aligned}\text{AIC}(\text{LIN}) &= \log[2/3] - 3 \\ &= -0.405 - 3 \\ &= -3.405\end{aligned}$$

There's nothing special about 2/3 here. If Pr(data | f(LIN)) is greater than $1/e^{12}$ (which is roughly equal to 0.00000614), LIN's AIC score will be greater than POLY-14's. In LIN the possibility of error is represented by a single adjustable parameter (e); this buys the model good fit-to-data while still keeping the model fairly simple. With POLY-14, you get perfect fit-to-data, but with a huge cost in complexity.

AIC raises a host of fascinating philosophical questions:

- What assumptions need to be true for AIC to provide an unbiased estimate of a model's predictive accuracy? In particular, does AIC depend for its justification on a principle of parsimony?
- There are other properties of an estimator that bear on whether it is "good" besides its being unbiased (e.g., its variance). How does AIC measure up with respect to those?
- Is a model's number of adjustable parameters a language-dependent quantity?
- How is it possible for a false model to be more predictively accurate than a true one?
- How is AIC related to the problem of induction and the grue problem?

The reader is encouraged to consult the literature on AIC already mentioned as a starting point. For present purposes, however, we ask the reader to recognize something more modest: models that over-fit the data at hand are apt to be predictively inaccurate, and one way to address that problem is to give some weight to how complex the model is (as measured by the number of adjustable parameters it contains). AIC provides a framework for making these intuitive ideas precise.

The example of POLY-14 and LIN can be modified slightly to illustrate how the quest for predictive accuracy differs from the quest for truth (or for probable truth). Add an error term to POLY-14, and call the resulting model POLY-14e. LIN entails POLY-14e; the former is a special case of the latter, since LIN can be obtained from POLY-14e by assigning zeroes to all but three of the adjustable parameters in that more complex model. Since LIN entails POLY-14e, but not conversely, LIN can't have the higher posterior probability, no matter what one's observations are. The point of interest is that LIN can still have the higher estimated predictive

accuracy; whether this is so depends on the data at hand. In addition, even if you know that POLY-14e is true and LIN is false, AIC may tell you to use LIN rather than POLY-14e to predict new data, again depending on the data you have. This example of nested models illustrates why AIC is nonBayesian; the concept of predictive accuracy it deploys is different from the concept of probable truth.[20]

The subject matters of LIN and POLY-14 are the nonmental variables x and y. What has that to do with the semantic content of a frog's belief state? The curve-fitting example involves quantitative variables x and y and models that contain adjustable parameters, but what are the quantitative variables and what are the adjustable parameters in models that attribute beliefs to organisms? This is the subject to which we now turn.

5 Causal models and the semantic contents of perceptual beliefs[21]

One of the uses that Fodor made of the disjunction problem (distinct from the uses described in footnote 3) was to criticize a theory of semantic content that he called "the Crude Causal Theory, [which] says, in effect, that a symbol expresses a property if it's nomologically necessary that all and only instances of the property cause tokenings of the symbol" (Fodor 1987, p. 100). Here we consider a different causal theory, which gives Fodor's crudity a probabilistic spin. It says that

> C_1: Neural state n means P precisely when P is a positive causal factor for the occurrence of n.

Positive causal factors don't need to necessitate their effects; they need only raise the probability of their effects when other causal factors are held fixed; see Hitchcock (2012) for discussion. This idea is broadly in accord with the interventionist account of causation presented by Woodward (2005). We don't claim that C_1 is an adequate theory of semantic content; rather, our hope is that lessons drawn from it concerning the status of hypotheses of semantic content that entail that some beliefs are false will carry over to more adequate theories.

C_1 is a connecting principle (thus the letter "C" in "C_1") that links hypotheses about meaning to causal models. We understand the "precisely when" to be stronger than the material biconditional. Think of C_1 as a theoretical postulate, one that is nomologically necessary if true at all. This interpretation of C_1 leads to a further connecting principle:

> C_2: AIC(n means P | data) = AIC(P causes n | data).

---

[20] Something similar is true in the context of "favoring" in the sense of the Law of Likelihood. If $Pr(O \mid H_1)$ is greater than $Pr(O \mid H_2)$, then O favors H_1 over H_2. It doesn't follow, though, that H_1 is probably true and H_2 is probably false.

[21] Here we adapt the analysis of how AIC applies to different causal models described by Forster and Sober (1994) and Sober (2015).

Keep in mind that our interest here is not in the truth of claims about what n means, but in the predictive accuracies of such claims.

With C_2 in hand, let's now consider several models for what causes frogs to go into neural state n. We won't (because no one can) consider all possible causal models for what makes frogs go into neural state n; rather, we'll consider a handful of causal models that involve FLY and BB, and some of their Boolean combinations. We'll begin with a few to give the reader a feeling for how this modeling exercise works, and then introduce a few more; these latter items will allow us to take up epistemological versions of the disjunction and distality problems.

Table 4: Pr(N | ±FLY & ±BB)		
	BB	**not-BB**
FLY	$x+f+s+i$	$x+f$
not-FLY	$x+s$	X

Each cell in Table 4 represents the probability of N (that neural state n occurs), conditional on what is going on in the frog's environment.[22] Using this table, the following causal models can be constructed:

Null: Neither flies nor beebees cause n (f=0, s=0, and i=0).

Flies Only: Flies cause n but beebees don't (f=α>0, s=0, and i=0).

Beebees Only: Beebees cause n but flies don't (f=0, s=β>0 and i=0).

Flies & Beebees Additive: Flies cause n and so do beebees (f=α>0, s=β>0, and i=0).

Flies & Beebees Non-Additive: Flies cause n and so do beebees (f=α>0, s=β>0, and i=γ≠0).

These models differ in their probabilistic implications. *Null* implies each of the following:

(1) $\Pr(N \mid FLY \,\&\, BB) - \Pr(N \mid \text{not-FLY} \,\&\, BB) =$

$\Pr(N \mid FLY \,\&\, \text{not-BB}) - \Pr(N \mid \text{not-FLY} \,\&\, \text{not-BB}) = 0$

[22] We earlier assumed that FLY and BB are mutually exclusive. We're now dropping that assumption – we assume that a visual scene can include a fly and a beebee too.

(2) $\Pr(N \mid FLY \,\&\, BB) - \Pr(N \mid FLY \,\&\, \text{not-}BB) =$

$\Pr(N \mid \text{not-}FLY \,\&\, BB) - \Pr(N \mid \text{not-}FLY \,\&\, \text{not-}BB) = 0$

Flies Only implies (2) but does not imply (1). In fact, it implies that (1) is false because it says that the following is true:

(3) $\Pr(N \mid FLY \,\&\, BB) - \Pr(N \mid \text{not-}FLY \,\&\, BB) =$

$\Pr(N \mid FLY \,\&\, \text{not-}BB) - \Pr(N \mid \text{not-}FLY \,\&\, \text{not-}BB) = \alpha > 0$

And so on for the other models.[23]

Our models differ in their probabilistic implications, but they also differ in their number of adjustable parameters. These are represented in Table 5. *Null* is the least complex and *Flies & Beebees Non-Additive* is the most. The other three models fall in between.

Table 5: Five causal models

Causal Models	**Parameters (adjustable, or set equal to zero)**			**Number of adjustable parameters**
Null	$f=0$	$s=0$	$i=0$	0
Flies Only	$f=\alpha$	$s=0$	$i=0$	1
Beebees Only	$f=0$	$s=\beta$	$i=0$	1
Flies & Beebees Additive	$f=\alpha$	$s=\beta$	$i=0$	2
Flies & Beebees Non-Additive	$f=\alpha$	$s=\beta$	$i=\gamma$	3

To apply AIC to these models, the first step is to use the frequency data represented in Table 6 to estimate the values of the adjustable parameters in each model. There are possible

[23] *Flies & Beebees Additive* says that the effect of the two causes on the probability of N when both are present is the sum of the effects of each cause on N when it is present and the other is absent. *Flies & Beebees Non-Additive* denies this because it introduces an interaction term *i* (whose value is stipulated to be nonzero).

data sets that the first four models will fit only imperfectly. For instance, *Flies Only* will fail to fit the data perfectly if $g_1 \neq g_2$ or if $g_3 \neq g_4$. In general, the more adjustable parameters, the better a model will fit the data. In fact, the most complex model on the list, *Flies & Beebees Non-Additive*, will fit the data perfectly, regardless of what the four sample frequencies turn out to be. However, that model buys its perfect fit at a cost: it is complex.

All of the models described so far, except for *Null*, will say that there are false beliefs in your data set, provided that $g_4 > 0$. In addition, *Flies Only* will say that mistakes were made if $g_3 > 0$, and *Beebees Only* has that implication if $g_2 > 0$.

Table 6: samplefreq(N | ±FLY & ±BB)		
	BB	**not-BB**
FLY	g_1	g_2
not-FLY	g_3	g_4

The use of AIC in the context of causal models has the obvious but important consequence that a causal model can't be evaluated when your sample frequencies don't allow maximum likelihood estimates to be made of one or more of its parameters. For example, a data set that describes how often neural state n occurs in the presence or absence of beebees and in the presence or absence of flies doesn't allow you to evaluate a model that says that little lumps of coal cause n. The data constrain the range of causal models you can test.

Consider now our second target question:

> Q_2: Let M_1 and M_2 be models of what causes neural state n in organism O. Are there situations where the data are such that (i) AIC(M_1 | data) > AIC(M_2 | data) and (ii) O has more false beliefs if M_1 is true than if M_2 is true?

This question can be answered by appeal to the following disjunctive model:

> *Flies-or-Beebees*: Flies-or-beebees cause n (f=α>0, s=α>0, and i=-α).

This model entails that the frog is utterly insensitive to the difference between flies and beebees in that:

(4) Pr(N | FLY & BB) = Pr(N | not-FLY & BB) = Pr(N | FLY & not-BB).

Flies Only and *Flies-or-Beebees* both have just one adjustable parameter, and so:

(5) AIC(*Flies Only* | data) > AIC(*Flies-or-Beebees* | data) iff log[Pr(data | f(*Flies Only*))] > log[Pr(data | f(*Flies-or-Beebees*))].

Clearly, there can be cases where the data on hand is such that the right-hand side of (5) holds – for example, cases where the data on hand is such that $g_1 = g_2 >> g_3 > g_4$. But, suppose, the frog has more false beliefs if *Flies Only* is true than if *Flies-or-Beebees*. Hence the answer to Q_2, as with the answer to Q_1, is affirmative.

Recall that the disjunction problem can be understood as the problem of answering the metaphysical question $DISJ_1$ (the question of what makes it the case that n's content is X as opposed to the logically weaker proposition X∨Y or the logically stronger proposition X&Z). We haven't tried to answer this question. Our point is that a model on which n's content is FLY can be better or worse in terms of expected predictive accuracy than a model on which n's content is the logically weaker proposition FLY-or-BB; which way the cards fall depends on the data.

As noted at the start of this paper, it sounds weird to say that a frog that can't tell a fly from a beebee could have the belief that a fly or a beebee is present. This makes a theory of content look crazy if it assigns this disjunctive belief to a nondiscrimating frog. However, this is quite tangential to the question of how misrepresentation is possible. Instead of the disjunctive model just described, consider one that works with the semantic content "that's a small dark object" and assume that in the frog's environment, small dark objects are either flies or beebees. That's a perfectly respectable model and AIC evaluates it in the same way we just evaluated *Flies-or-Beebees*.

We turn now to our third, and final, target question:

Q_3: Let M_1 and M_2 be models of what causes neural state n in organism O. Are there situations where the data are such that (i) AIC(M_1 | data) > AIC(M_2 | data), (ii) n is about something external to O's body if M_1 is true, and (iii) n is about O's retinal states if M_2 is true?

This question is similar to but distinct from the metaphysical question $DIST_1$ (the question of what makes it the case that n's content is X as opposed to the more causally distal proposition C_d, and the more causally proximal proposition C_p). We won't try to answer the latter question. Our aim, rather, is to answer Q_3.

Consider the following retinal model:

Retinal States: r_1, r_2, …, and r_{10} cause n, where r_1, r_2, …, and r_{10} are incompatible (but not jointly exhaustive) retinal states, and $Pr(N \mid R_i) - Pr(N) = \alpha_i > 0$ for i = 1, 2, …, 10.

It can easily turn out that:

(6) AIC(*Flies Only* | data) > AIC(*Retinal States* | data).

Suppose, for example, that Pr(data | f(*Flies Only*)) equals 1/10 and Pr(data | f(*Retinal States*)) equals 9/10. Then AIC(*Flies Only*) is roughly equal to –3.303 whereas AIC(*Retinal States*) is roughly equal to –10.105. The key here is that though *Flies Only* is worse than *Retinal States* in terms of fit-to-data, the former is much better than the latter in terms of simplicity; the former has 1 adjustable parameter (α) whereas the latter has 10 (α_1, α_2, …, α_{10}).[24] Hence the answer to Q_3, as with the answers to Q_1 and Q_2, is affirmative.

The present example comes from Dretske (1981, Ch. 6), but his solution of the distality problem is different. Suppose a fly can cause different retinal states in a frog, each of which can trigger the same neural state n. Dretske notes that n may indicate that a fly is present without indicating which of those k retinal states has occurred. He suggests that this explains why the frog's neural state n has the experiential content FLY. If we apply Dretske's analysis to the topic of perceptual belief, it encounters a difficulty. By assumption, neural state n indicates that a fly is present, but does not indicate which retinal state is occurring. However, the neural state does indicate that a *disjunction* is true, where each disjunct says that one or another of those k retinal states is occurring. Thus, the indication relation, by itself, does not solve the distality problem. Dretske recognizes this point and invokes "specificity" to solve it. In later work, he suggests a different solution − functional considerations (Dretske 1986, 1988). The frog sees that a fly is present, not that a retinal disjunction is true, Dretske says, because the function of neural state n is to indicate that flies are present, not to indicate that disjunctive retinal events are occurring. We, in contrast, are not proposing a theory that tells you why the frog's neural state n has the experiential content FLY, nor are we invoking functional considerations. Rather, we are describing how hypotheses about the causes of that neural state may be evaluated for their predictive accuracies.

We so far have described how scientists can evaluate meaning hypotheses for their predictive accuracies (given connecting principles C_1 and C_2). We have talked about using observations of a current frog to evaluate models about the causation of a neural state, but have said nothing about the evolution of the frog in question. Even so, predictive accuracy has a role

---

[24] This argument does not carry over to retinal causal models like the following:

> *Retinal States**: r_1, r_2, …, and r_{10} cause n, where r_1, r_2, …, and r_{10} are incompatible (but not jointly exhaustive) retinal states, and each is such that $Pr(N \mid R_i) - Pr(N) = \alpha > 0$ for i = 1, 2, …, 10.

Here there is just 1 adjustable parameter (α). If Pr(data | f(*Flies Only*)) equals 1/10 and Pr(data | f(*Retinal States**)) equals 9/10, then AIC(*Flies Only*) is roughly equal to –3.303 but AIC(*Retinal States**) is roughly equal to –1.105. Perceptual phenomena like size and shape constancy (Sternberg 2006) suggest that models like *Retinal States* may be more predictively accurate than models like *Retinal States**. This is because the different retinal states that increase the probability of a given neural state often increase that probability by significantly different amounts.

to play in evolutionary considerations. We conjecture that frog brains create models that allow for the possibility of misrepresentation for the same reason that human theorists create models that allow for the possibility of error.[25] In both cases, models that fit the present data imperfectly may be more predictively accurate than models that fit that data perfectly. Frogs, like scientists, need to deploy models that optimally balance fit-to-data against simplicity. AIC provides a formula for comparing how well different models succeed in achieving that trade-off.

We are not saying that frogs are mathematical statisticians. However, we do think that the evolutionary argument in Section 3 of this paper can be supplemented with the following conjecture. Natural selection often favors organisms that use models that are more predictively accurate over organisms that use models that are less so, and organisms achieve predictive accuracy by balancing fit-to-data against simplicity. The result may be a selective advantage that goes to organisms that formulate false beliefs, frogs and human beings alike.[26]

6 Concluding comments

The two-part epistemology proposed in this paper is half Bayesian. In Sections 2 and 3, we assigned probabilities to states of the world and calculated expected utilitities, just as Bayesians would wish. However, in Sections 4 and 5, we used AIC as a tool for estimating predictive accuracies, but AIC does not estimate how probable it is that a model's predictions will have a given degree of accuracy. AIC is nonBayesian.[27]

We began by considering three propositions:

(a) The organisms in a population work with the same set of propositions.

(b) When a proposition is true, organisms do better by believing it than by disbelieving it.

---

[25] Here we don't have in mind propositions such as FLY, BB, FLY-or-BB, and not-FLY&not-BB from Section 3. These propositions have no adjustable parameters and thus are *not* models strictly speaking. AIC entails that they are to be evaluated just by their fit-to-data. In contrast, consider the example depicted in Figure 2, and suppose that a frog needs to find a predictively accurate curve, given the data points it has observed.

[26] AIC thus provides a possible explanation for why human beings use "fast and frugal" heuristics. They deploy simple models that require less information to apply and are better predictors than highly complex models that require more information to apply (Gigerenzer and Goldstein 1996). See Forster (1999) for discussion.

[27] There is a Bayesian criterion for model selection called BIC, which, like AIC, takes account of both fit-to-data and complexity as measured by number of adjustable parameters (Schwarz 1978). For discussion of how AIC and BIC are related to each other, see Sober (2015).

(c) The payoffs to organisms that believe the same proposition in the same state of the world are equal.

We argued that when these three conditions are satisfied, selection will favor increasing the world-to-head reliability of belief formation mechanisms. The fittest possible organism will have error probabilities of zero.

Matters change when different organisms work with different sets of propositions. Then, a risk-taking organism can be fitter than a risk-free organism, even when (b) and (c) are true. In particular, it is the expansion of representational capacities – e.g., being able to discriminate between flies and beebees, and behaving differently as a result − that opens the door for natural selection to move the population towards increased risk-taking in belief formation. This argument does not explain why natural selection should produce organisms that have beliefs. Rather, our question concerns the sort of belief-formation mechanisms that organisms should have, given that they have beliefs.

A simple model that contains an error term will often fit the data less well than a more complex model that has no error term and fits the data perfectly. Even so, the former model may have the better estimated predictive accuracy (depending on the data). AIC explains why organisms that use models that fit the data imperfectly are often more accurate predictors than organisms that use models that fit the data perfectly. The organisms in question can be frogs or scientists.

Although our discussion of risk-taking and risk-avoidance in Sections 2 and 3 was thoroughly evolutionary, our discussion of over-fitting and predictive accuracy in Sections 4 and 5 was not, until the very end of that section. AIC allows scientists to use data on an organism's behavior to model what the organism believes without the scientist needing to think about how the organism evolved. This point is compatible with AIC's having an evolutionary significance, as we explained.

The AIC framework also helps explain why a simple model concerning a single distal cause of a given effect may be more predictively accurate than a more complex model concerning multiple proximate causes that that effect may have. If connecting principles C_1 and C_2 are true, hypotheses about the meaning of a neural state may differ in their predictive accuracies. Scientists may be better able to predict the neural state's occurrence by the simpler distal causal model than by the more complex model of proximate causes. We addressed epistemic versions of the disjunction and distality problems using AIC, but AIC applies to other kinds of meaning hypothesis as well.

In closing, we want to comment on how our analysis comes in contact with Donald Davidson's much-discussed "principle of charity." The principle has been formulated in different ways by different commentators, but one simple (perhaps oversimple) formulation of Davidson's idea deserves comment. It is:

D: We should interpret speakers as holding true beliefs (true by our lights at least) as much as possible.

This isn't a bad gloss of Davidson's (1970) statement that "in our need to make him make sense we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights it goes without saying)." The "as much as possible" rider in proposition D concedes that sometimes it may not be possible to interpret 100% of a speaker's utterances as true. However, D entails that when one possible interpretation says that all the speaker's utterances are true while another says that only 90% are, the former is better.

The topic of our paper isn't the interpretation of utterances but the attribution of beliefs. We've assumed that this can be done in connection with organisms (e.g., frogs) that do not speak a language as well as in connection with organisms that do. Davidson (1975, 1982) famously defended the thesis that thought without language is impossible, so he would reject our froggy assumption out of hand. However, our argument also applies to organisms that do speak a language, and here we and Davidson disagree.

References

Adams, Fred and Aizawa, Kenneth. (2017): "Causal Theories of Mental Content," in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Summer 2017), URL = <https://plato.stanford.edu/archives/sum2017/entries/content-causal/>.

Akaike, Hirotugu (1973): "Information Theory as an Extension of the Maximum Likelihood Principle," in B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory* (Budapest: Akademiai Kiado, pp. 267-281).

Axelrod, Robert (1980): *The Evolution of Cooperation.* New York: Basic Books.

Burnham, Kenneth and Anderson, David. (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York, NY: Springer-Verlag.

Clark, Andy (1993): "Mice, Shrews, and Misrepresentation," *The Journal of Philosophy* 90: 290-310.

Davidson, Donald (1970): "Mental Events." Reprinted in Davidson 1980, *Essays on Actions and Events*.

Davidson, Donald (1975): "Thought and Talk." Reprinted in Davidson 1984, *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davidson, Donald (1982): "Rational Animals." Reprinted in Davidson 2001, *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press.

Dretske, Fred (1981): *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.

Dretske, Fred (1986): "Misrepresentation," in R. Bogdan (ed.), *Belief* (Oxford: Oxford University Press, pp. 17-36).

Dretske, Fred (1988): *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: MIT Press.

Field, Hartry (1990): "'Narrow' Aspects of Intentionality and the Information-Theoretic Approach to Content," in E. Villanueva (ed.), *Information, Semantics, and Epistemology* (Oxford: Blackwell, pp. 102-116).

Fodor, Jerry (1984): "Semantics, Wisconsin Style," *Synthese* 59: 231-250.

Fodor, Jerry (1987): *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Forster, Malcolm (1999): "How Do Simple Rules Fit to Reality in a Complex World?" *Minds and Machines* 9: 543-564.

Forster, Malcolm and Sober, Elliott (1994): "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science* 45: 1-36.

Gendler, Tamar (2008): "Alief and Belief," *Journal of Philosophy* 105: 634-663.

Gibson, Martha (1996): "Asymmetric Dependencies, Ideal Conditions, and Meaning," *Philosophical Psychology* 9: 235-259.

Gigerenzer, Gerd and Goldstein, Daniel (1996): "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review* 103: 650-669.

Godfrey-Smith, Peter (1991): "Signal, Decision, Action," *Journal of Philosophy* 88: 709-722.

Godfrey-Smith, Peter (1996): *Complexity and the Function of Mind in Nature.* Cambridge, UK: Cambridge University Press.

Grice, Paul (1957): "Meaning," *The Philosophical Review* 66: 377-388.

Hitchcock, Christopher (2012): "Probabilistic Causation," in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (winter ed.). URL = <http://plato.stanford.edu/entries/causation-probabilistic/>.

James, William (1897): "The Will to Believe," in *The Will to Believe and Other Popular Essays in Philosophy* (Cambridge, MA: Harvard University Press, 1979, pp. 1-31).

Krebs, John and Davies, Nicholas (1981): *An Introduction to Behavioral Ecology*. Oxford: Sinauer.

Lettvin, Jerome, Maturana, Humberto, McCulloch, Warren and Pitts, Walter (1959): "What the Frog's Eye Tells the Frog's Brain," *Proceedings of the Institute of Radio Engineers* 47: 1940-1951.

McLaughlin, Brian. (1987): "What is Wrong with Correlational Psychosemantics?" *Synthese* 70: 271-286.

McKay, Ryan and Dennett, Daniel (2009): "The Evolution of Misbelief," *Behavioral and Brain Sciences* 32: 493-510.

Neander, Karen (2017): *A Mark of the Mental: In Defense of Informational Teleosemantics* Cambridge, MA: MIT Press.

Neander, Karen (2018): "Teleological Theories of Mental Content," in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2018), URL = <https://plato.stanford.edu/archives/spr2018/entries/content-teleological/>.

Schwarz, Gideon (1978): "Estimating the Dimension of a Model," *Annals of Statistics* 6: 461-464.

Shapiro, Lawrence (1997): "The Nature of Nature: Rethinking Naturalistic Theories of Intentionality," *Philosophical Psychology* 10: 309-322.

Sober, Elliott. (1980): "Evolution, Population Thinking, and Essentialism," *Philosophy of Science* 47: 350-383.

Sober, Elliott (1994): "The Adaptive Advantage of Learning and *A Priori* Prejudice," in *From a Biological Point of View.* Cambridge: Cambridge University Press.

Sober, Elliott (1997): "Is the Mind an Adaptation for Coping with Environmental Complexity? A Review of Peter Godfrey-Smith's *Complexity and the Function of Mind in Nature*." *Biology and Philosophy* 12: 539-550.

Sober, Elliott. (2015): *Ockham's Razor – A User's Manual.* Cambridge: Cambridge University Press.

Stampe, Dennis (1977): "Toward a Causal Theory of Linguistic Representation," in P. French, H.K. Wettstein, and T.E. Uehling (eds.), *Midwest Studies in Philosophy*, vol. 2 (Minneapolis: University of Minnesota Press, pp. 42-63).

Stephens, Christopher (2001): "When is it Selectively Advantageous to Have True Beliefs? Sandwiching the Better-Safe-than-Sorry Argument," *Philosophical Studies* 105: 161-189.

Sterelny, Kim (1990): *The Representational Theory of Mind: An Introduction*. Oxford: Blackwell.

Sterelny, Kim (2003): *Thought in a Hostile World*. Oxford: Blackwell.

Sternberg, Robert (2006): *Cognitive Psychology*. Belmont, CA: Wadsworth, Cengage Learning.

Woodward, James (2005): *Making Things Happen.* Oxford: Oxford University Press.